VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

KONVERZE HLASU

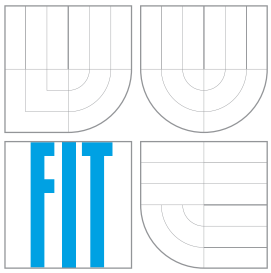
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

IVAN SCHWARZ

BRNO 2010

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

KONVERZE HLASU

VOICE CONVERSION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — IVAN SCHWARZ
AUTHOR

VEDOUCÍ PRÁCE — doc. Dr. Ing JAN ČERNOCKÝ
SUPERVISOR

BRNO 2010

Abstrakt

Práce je věnována tvorbě systému pro konverzi hlasu. Metodám, jež hlas jednoho člověka upraví tak, aby ho bylo možné zaměnit za hlas člověka jiného. V první části je popsán harmonický a šumový model (HNM), který se stará o analýzu a syntézu signálů. Druhá část se zabývá metodami konverze. Nejprve jsou uvedeny prozodické změny a následně i možnosti modifikace spektrální obálky, zejména použití konverzních matic. Stručně je zde vysvětlena metoda dynamického borcení času (DTW) a metoda kódování pomocí lineární predikce (LPC). V poslední části je uveden způsob implementace, popsán průběh testování a jsou diskutovány dosažené výsledky. V závěru jsou nastíněny možnosti dalšího vývoje.

Abstract

Thesis is dedicated to the making of a system for voice conversion. To methods, which alter voice of one person in a way, that it could be possible for listener to mislead it for someone elses voice. In the first part, Harmonic plus Noise Model (HNM) is described. Signal analysis and synthesis are its main purposes. Methods of voice conversion are considered in the second part. Prosodic modifications are introduced at first and then modification of a spectral envelope is discussed (Especially aplication of conversion matrices). Dynamic Time Warping (DTW) and Linear Prediction Coding (LPC) methods are explained briefly. In last section, implementation process is described and achived results are discussed. Ways of further development are suggested in summary.

Klíčová slova

hlas, řeč, konverze hlasu, modifikace řečníka, harmonický a šumový model, prozodie, spektrální obálka, dynamické borcení času, lineární predikce

Keywords

voice, speech, voice conversion, speaker modification, harmonic plus noise model, prosody, spectral envelope, dynamic time warping, linear prediction

Citace

Ivan Schwarz: Konverze hlasu, bakalářská práce, Brno, FIT VUT v Brně, 2010

Konverze hlasu

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením doc. Dr. Ing. Jana Černockého.
Další informace mi poskytli Ing. Igor Szöke a Ing. Yannis Stylianou, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Ivan Schwarz
19. května 2010

Poděkování

Rád bych poděkoval vedoucímu práce Janu Černockému za čas strávený pravidelnými konzultacemi a trpělivost při osvětlování temných oblastí zpracování řeči. Děkuji Igoru Szökemu za poskytnutou implementaci harmonického a šumového modelu a informace k této implementaci se vztahující. Také děkuji Yannisovi Stylianou za aktuální informace související s tématem Konverze hlasu a Janu Holasovi za pomoc při hodnocení dosažených výsledků.

Obsah

Kapitola 1

Úvod

Lidský hlas je zřejmě nejpřirozenější prostředek ke komunikaci. Kromě své nedůležitější funkce, přenosu myšlenky, obsahuje spoustu dalších zajímavých informací. Mezi ně patří například emoční stav mluvčího, může mnoho napovědět o jeho zdravotním stavu, či dokonce fyzických vlastnostech. Všechny tyto součásti napomáhají k tomu, aby byl každý hlas do jisté míry jedinečný a my tak mohli s určitou pravděpodobností rozpoznat jeho vlastníka.

Konverze hlasu se věnuje metodám, které hlas jednoho člověka (zdrojového mluvčího) upraví do takové míry, že ho bude možné zaměnit za hlas člověka odlišného (cílového mluvčího). Abychom takto mohli hlas modifikovat, je nutné nejprve získat dostatečně dlouhé vzorky řeči obou subjektů, analyzovat je a určit tak potřebné řečové parametry. Tyto pak vzájemně porovnáme a vypočítáme koeficienty, které nakonec poslouží k samotné modifikaci řečových parametrů zdrojového mluvčího tak, aby po následné syntéze odpovídaly řeči cílového mluvčího.

Základem pro práci s řečí jsou tedy dvě fáze. Její analýza a její syntéza. K jejich uskutečnění existuje několik metod. V tomto dokumentu bude popsán Harmonický a šumový model (Dále označovaný pouze jako HNM - Harmonic plus Noise Model) tak, jak byl prezentovaný v [5] a následně upravený a implementovaný v [6].

Konverzi hlasu rozdělíme na několik částí. Nejprve provedeme změny souhrnně označované jako prozodické. Jedná se o modifikaci rytmu, melodie a hlasitosti řeči. Poté se zaměříme na spektrální obálku signálu, její podíl na identifikaci mluvčího posluchačem a možnosti její úpravy.

Vzhledem k tomu, že některé skupiny fonémů nesou více informací potřebných k rozpoznání řečníka než jiné, bude využit fonémový rozpoznávač vyvíjený skupinou Speech@FIT a popsaný v [4].

Představíme si výsledky jednotlivých modifikací i jejich kombinací. Zvážíme přínos fonémového rozpoznávače. Budeme hodnotit rychlost a úspěšnost. Přihlédneme ke kvalitě nahrávek a k tomu, zda byla k dispozici identická nahrávka zdrojového i cílového mluvčího či nikoliv. Posuzovat budeme subjektivně poslechem.

Kapitola 2

Harmonický a šumový model

Reálný řečový signál se skládá ze znělé části a z části šumové. Jak je vysvětleno v [5], k modelování znělé části se ukázaly jako velmi efektivní sinusoidní modely. Tyto ale nepracují s šumovou složkou a jejich použití při modifikaci signálu, zejména při změně rytmu či melodie, je velice problematické. Proto se k tomuto účelu obvykle využívají modely hybridní, které daný signál rozdělí na složku deterministickou a složku stochastickou. Na každou z nich jsou pak užity rozdílné metody zpracování a jak ukázaly experimenty v [5] i v [6], tento způsob vede k výraznému zlepšení kvality modifikovaného syntetizovaného signálu.

2.1 Popis harmonického a šumového modelu

Rozdělení řečového signálu se provádí na základě jeho frekvenčního spektra (obr. 2.1d), které s využitím časově proměnlivé hranice maximální znělé frekvence (jež je jedním z důležitých parametrů HNM, o kterém si více povíme v následující podkapitole) vymezuje spodní, převážně harmonickou část (obr. 2.1b) a vrchní, převážně šumovou část (obr. 2.1c). Obě části se totiž prolínají. Harmonickou část h(t) vyjádříme následujícím vztahem.

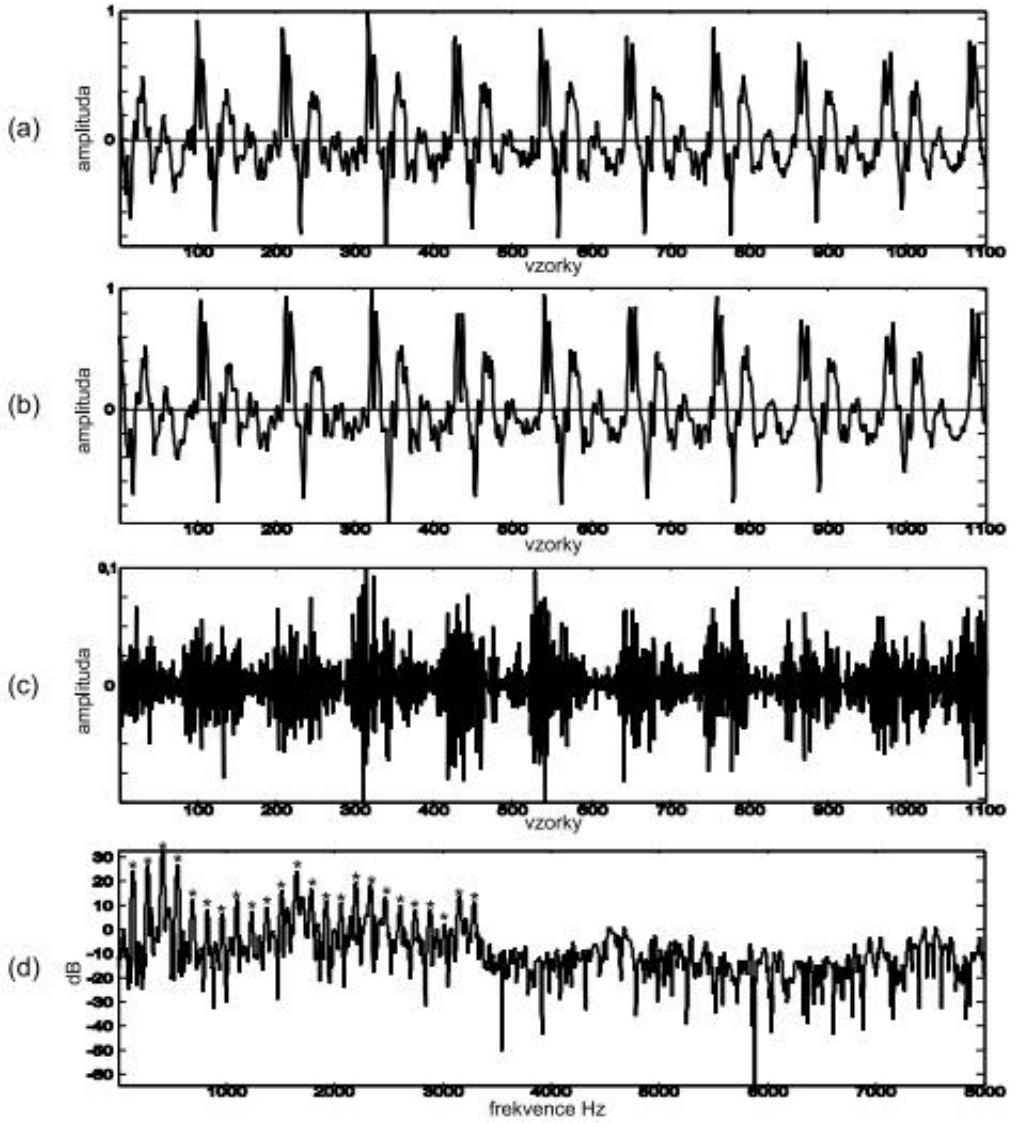
$$h(t) = \sum_{k=1}^{K(t)} a_k(t) \cos \phi_k(t), \tag{2.1}$$

kde a$_k(t)$ je amplituda a $\phi_k(t)$ je fáze k-té harmonické složky v čase t. K(t) je časově proměnný parametr a ukazuje počet harmonických násobků frekvence základního tónu v harmonické části.

$$K(t) = \frac{F_m(t)}{F_0(t)} \tag{2.2}$$

Pro popis šumové složky existuje více možností. V přístupu k HNM, prezentovaném v [6], jehož implementaci budeme využívat, se autor přiklání k metodě, která nevyužívá OLA[1] ani filtrování a parametrizuje šumovou část se stejným principem, jako tu harmonickou. Jelikož zde ale neexistuje perioda základního tónu, je zvolena pevně. Výsledný signál pak získáme jednoduše sečtením harmonické a šumové složky.

[1] OverLap Add method



Obrázek 2.1: Obrázek ukazuje jednotlivé složky signálu samohlásky "e". celkový signál s(t) je na a). Ten jsme dále rozdělili frekvencí 4500 Hz na část b) obsahující převážně harmonickou složku h(t) a na část c) obsahující převážně šumovou složku n(t). V d) je zobrazeno spektrum signálu a) v dB. Můžeme si všimnout, že v části $0-3200$ Hz jsou patrné oblasti harmonických násobků frekvence základního tónu hlasu. Obrázek i s popisem byl převzat z [6]

2.2 Modelování signálu pomocí HNM

Konverze hlasu se obecně skládá ze tří fází. Z fáze *analýzy*, z fáze *modifikace řečových parametrů* a nakonec z fáze *syntézy*. K první a poslední fázi využijeme právě výše zmiňovaný HNM.

Fáze analýzy bývá časově náročná, protože se při ní získávají všechny důležité řečové parametry. V prvním kroku signál ustředíme, získáme z něj základní parametry, jako je frek-

vence základního tónu $F_0(t)$, maximální znělá frekvence $F_m(t)$ a určíme znělost/neznělost částí signálu. V druhém kroku vypočítáme amplitudy $A_k(t)$, fáze $\phi_k(t)$ a jejich spektrální a časové obálky. Tato fáze se provádí pouze jednou.

Fáze syntézy je rychlejší a provádí se vždy, když vytváříme nový, modifikovaný signál, po úpravě řečových parametrů za pomoci příslušných koeficientů.

2.3 Příprava signálu pro určení parametrů HNM

Než začneme s analýzou signálu a s určováním jeho parametrů, je vhodné si signál předpřipravit. Nejprve odstraníme případné chyby způsobené nahráváním signálu jeho tzv. *ustředěním*. Toho docílíme odečtením stejnosměrné složky například tak, jak je ukázáno v následujícím vztahu:

$$s^{'}(n) = s(n) - \frac{\sum_{x=1}^{N} s(x)}{N}, \tag{2.3}$$

kde N značí počet prvků vstupního signálu.

Aby se nám lépe určovala frekvence základního tónu a další důležité parametry, signál rozdělíme na rámce. Signál pak bude kratší a lépe analyzovatelný. Podle [5] je ideální délka rámce 2 - 3 hlasivkové periody. Nyní určíme periodu hlasivkových kmitů. Jelikož je ale v každém rámci period více, budeme je překrývat. V ideálním případě tak, aby každý následující byl vždy posunut o délku periody hlasivkového kmitu. Takto s vysokou přesností získáme skutečnou délku každé periody.

2.4 Základní parametry harmonického a šumového modelu

Prvním důležitým parametrem HNM je *základní tón* řeči, který má hlavní podíl při modifikaci melodie řečníka. Tento parametr se v okamžiku začátku každé periody hlasivkového kmitu může změnit. V průběhu této periody ovšem zůstává konstantní.

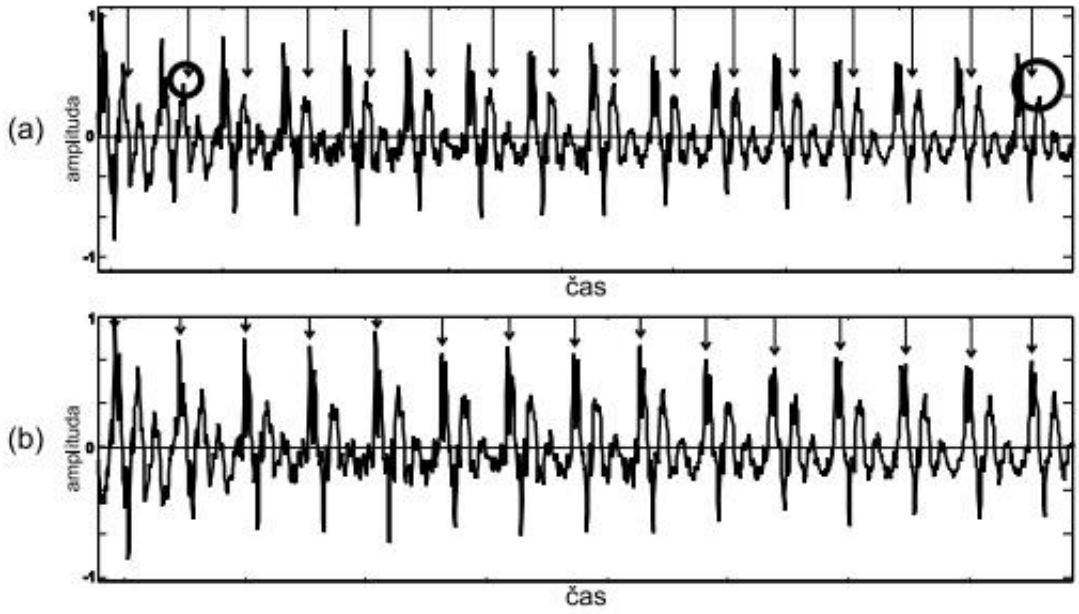
Nejprve se pokusíme odhadnout hrubý průběh této funkce. Existuje pro to několik metod, z nichž autor naší implementace vybral metodu autokorelační, která je rychlá a jednoduchá a jejíž spolehlivost ještě zvýšil dalšími úpravami, popsanými v [6]. Pro periodické signály je taktéž periodická a je sudá.

Jelikož jsme již určili alespoň přibližně průběh základního tónu a víme, že každá znělá řeč takový tón musí obsahovat, můžeme toho nyní využít k rozhodování, který rámec je znělý a který nikoliv. Pomocí následujícího vztahu vypočteme energii pro každý rámec.

$$E = \sum_{n=1}^{N} s^{'}(n)^2 \tag{2.4}$$

Nyní použijeme prahování s prahem $0{,}05\,E_{max}$. Pokud tedy energie jednotlivých rámců bude menší než 5 % energie celého signálu, budeme daný rámec považovat za neznělý.

Aby autor [6] zpřesnil určení základního tónu a zároveň přitom získal *okamžiky hlasivkové periody* t_a^i (kde i značí, s kolikátým okamžikem od začátku signálu pracujeme a a nám říká, že jde o okamžik z fáze analýzy), rozhodl se signál procházet po dříve odhadnutých periodách s tím, že se v každé pokusil najít maximum. Čím dále bylo nalezené maximum od předpokládaného okamžiku hlasivkové periody t_a^i, tím více bylo penalizováno. Tyto maxima byla následně prohlášena za okamžiky hlasivkové periody. Vliv tohoto kroku na přesnost odhadu základního tónu je vidět na obrázku 2.2.



Obrázek 2.2: Porovnání a) hrubého odhadu periody základního tónu znělé části signálu a b) zpřesněného odhadu na základě maxim blízkých k místu hrubého odhadu. I když je hrubý odhad a) relativně přesný, lze si všimnout posunu značek vůči signálu. V b) jsou tyto značky (t_a^i) již zcela synchronní s periodou základního tónu. Obrázek i s popisem byl převzat z [6]

Posledním důležitým parametrem HNM je *maximální znělá frekvence*. Tu budeme určovat pro všechny okamžiky hlasivkové periody. Toho bylo docíleno metodou popsanou v [5], pomocí detekce výrazných laloků ve spektru, jež jsou tvořeny harmonickým násobkem frekvence základního tónu. Maximální znělá frekvence je v čase velice proměnlivá. Vzhledem k tomu, že podle [6] ztrácí člověk schopnost vnímat fázi 4 - 5 kHz a vyšší a hodnoty nižší než 1 - 2 kHz snižují kvalitu syntézy harmonické složky, nemá smysl maximální znělou frekvenci nastavovat za tyto meze.

2.5 Amplitudy a fáze

Harmonická složka se vyskytuje ve znělých částech signálu v intervalu mezi F_0 a F_m. Jak jsme si už dříve řekli, k jejímu popisu využijeme sinusoidní model. Abychom ho ovšem mohli použít, potřebujeme určit amplitudy a fáze k-tého harmonického násobku základní frekvence pro všechny okamžiky hlasivkové periody. Toho docílíme s využitím komplexních exponenciál.

Kromě samotných amplitud a fází je důležité vytvořit i jejich obálky. Ty mají velký význam při získávání interpolovaných parametrů pro syntetizované okamžiky hlasivkových period (t_s^i). Obálku amplitud vypočítáme za pomocí diskrétního cepstra. Obálka průběhu fáze bude lineární. Zde je potřeba, aby její časový i frekvenční průběh byl co nejhladší. Přesnější postupy je možné nalézt v [6].

V šumové složce ponecháváme pouze amplitudy, fáze volíme náhodně. Při syntéze tak

do signálu zavedeme určitou stochastičnost. Jak už bylo řečeno v sekci 2.1, musíme zde volit periodu základního tónu pevně. Autor [6] použil hodnotu 10 ms. Maximální hodnota frekvence je rovna polovině vzorkovací frekvence a minimální frekvence je rovna maximální znělé frekvenci. Pokud znělá frekvence neexistuje, pak použijeme frekvenci základního tónu.

2.6 Fáze syntézy

Celkový signál se skládá z harmonické složky a z šumové složky. Při syntéze tyto dvě jednoduše sečteme.

Při syntéze harmonické části víceméně použijeme opačný postup než v analýze. Sečteme všechny harmonické složky s příslušnou fází a amplitudou. Jelikož ale některé zjednodušující opatření vedla k nespojitostem mezi jednotlivými okénky, bude nutné mezi nimi provést interpolaci.

Vzhledem k tomu, že v implementaci HNM podle [6] je popis harmonické složky stejný i pro složku šumovou, bude i syntéza stejná.

Kapitola 3

Konverze hlasu

Při návrhu systému pro konverzi hlasu je výhodné si uvědomit, jak člověk vnímá cizí hlas a podle čeho určuje identitu řečníka, kterému tento hlas patří. Tyto mechanismy můžeme metaforicky rozdělit na dvě skupiny. První tvoří "softwarové" mechanismy, ke kterým patří sémantické a lingvistické aspekty řeči. Tedy slovní zásoba, typický osobitý způsob vyslovování, dialekt, atp. Druhá skupina se skládá z mechanismů "hardwarových", týkajících se akustické stránky řeči. Jde spíše o samotný zvuk hlasu, který slyšíme. "Softwarové" mechanismy sice často obsahují důležitější informace pro identifikaci řečníka, ale jejich formální popis a stejně tak extrahování z řečového signálu, by bylo velmi obtížné. Naopak "hardwarové" mechanismy lze, jak jsme si ukázali v minulé kapitole, popsat mnohem snáze, řečovými parametry.

Z průzkumu popsaného v [5] vyplývá, že na jedinečnost hlasu má vliv kombinace několika různých faktorů. Jejich význam se z řečníka na řečníka liší. Ukázalo se ale, že velmi důležitou složkou individuality hlasu jsou základní tón a rytmus řeči, tedy prozodické parametry. Podstatnou část informací k identifikaci řečníka nese také jeho řečové spektrum. Pro odlišení jednotlivých řečníků se ukázaly jako obzvlášť užitečné charakteristické stopy, jež ve spektru zanechávají nosovky a samohlásky.

3.1 Modifikace prozodie

Analýza pomocí HNM nám připravila výbornou startovací pozici pro změnu prozodických vlastností signálu. Mezi nejdůležitější změny patří *modifikace rytmu*, *modifikace melodie* a *modifikace hlasitosti*, případně jejich kombinace. Takto budeme řečový signál alternovat poté, co jsme ho analyzovali a předtím, než ho budeme syntetizovat.

Abychom věděli, do jaké míry máme upravovat jednotlivé prozodické vlastnosti, musíme se zaměřit na jejich správné určení u obou řečníků a zjistit jak se liší.

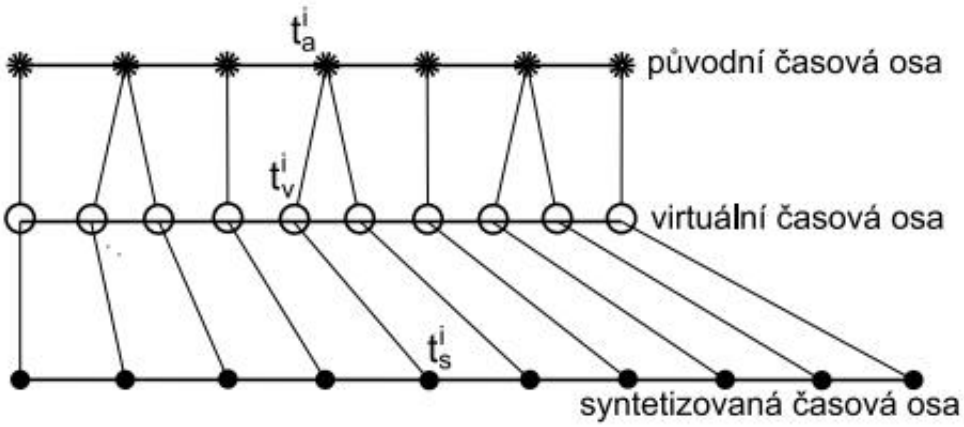
3.1.1 Modifikace rytmu

Modifikací rytmu rozumíme zrychlení nebo zpomalení tempa řeči aniž bychom ovlivnili její frekvenční průběh. Signál je rozdělen na rámce a okamžiky hlasivkové periody jsou již také určené. Za pomocí koeficientu $\beta(t)$, který bude určovat míru zrychlení/zpomalení můžeme funkci pro modifikaci rytmu D(t) vyjádřit takto:

$$D(t) = D(t_a^i) + \beta(t - t_a^i) \quad \text{pro} \quad t \in \ < t_a^i, t_a^{i+1} >, \tag{3.1}$$

kde $D(t_a^1) = 0$.

Jak již víme, při této modifikaci nelze připustit změnu periody základního tónu. Vzdálenost mezi t_s^i a t_s^{i+1} by měla být stejná, jako vzdálenost mezi t_a^i a t_a^{i+1}. K tomu nám napomůže vytvoření pomocné virtuální osy, stejné, jako je časová osa reálná. Na tuto osu umístíme virtuální okamžiky, které budou odpovídat syntetizovaným okamžikům. Vztahy mezi analyzovanými, virtuálními a syntetizovanými okamžiky hlasivkových period jsou znázorněné na obr. 3.1, který byl převzat z [6].



Obrázek 3.1: Vztahy mezi okamžiky t_a^i, t_v^i, t_s^i při modifikaci rytmu s koeficientem $\beta(t) = 1.5$.

Pokud pak chceme měnit rytmus řeči zdrojového mluvčího tak, aby odpovídal mluvčímu cílovému, změříme délku obou signálů, porovnáme je a koeficientu $\beta(t)$ přiřadíme příslušnou hodnotu. Je vhodné ignorovat často přítomné ticho na začátku a konci nahrávky a pracovat pouze se samotnou řečí.

3.1.2 Modifikace melodie

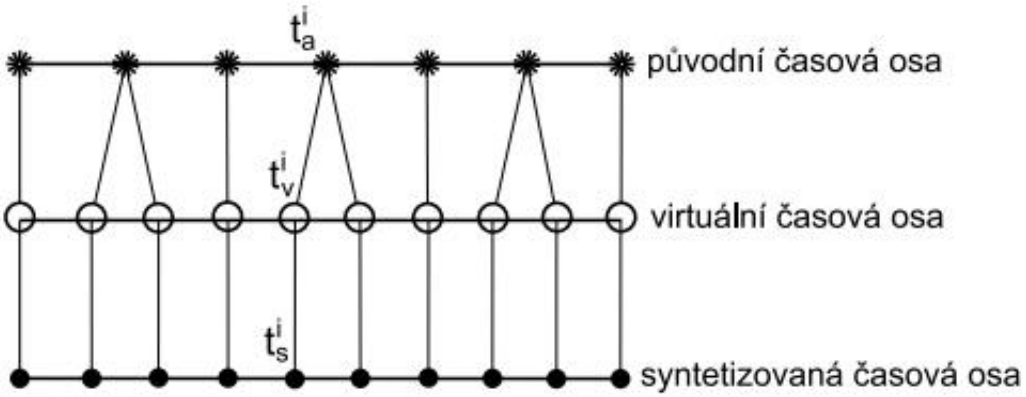
Modifikace melodie naopak spočívá v úpravě frekvenční oblasti signálu. Zde se snažíme, aby se nijak neměnil časový průběh. Jde zejména o úpravu periody P základního tónu řeči pomocí volitelného koeficientu $\alpha(t)$. Ten nadefinujeme tak, aby byl částech (vždy po dobu trvání okamžiku hlasivkové periody) konstantní.

$$\alpha(t) = \alpha(t_a^i) \quad \text{pro} \quad t \in \ < t_a^i, t_a^{i+1} > \tag{3.2}$$

Modifikaci vykonáváme postupně, pro všechny znělé okamžiky hlasivkové periody. Na obrázku 3.2 vidíme mapování analyzačních okamžiků t_a^i na syntetizační okamžiky t_s^i.

$$P'(t_s^i) = \frac{P(t_a^i)}{\alpha(t_a^i)} \tag{3.3}$$

Konverze mezi zdrojovým a cílovým řečníkem pak provádíme na základě porovnání jejich základních tónů. U obou mluvčích tudíž projdeme všechny okamžiky hlasivkové periody a u těch, které budou znělé, si poznamenáme výši základního tónu. Z nasbíraných hodnot určíme aritmetický průměr. Průměrné základní tóny mezi sebou porovnáme a výsledek dosadíme do koeficientu $\alpha(t)$.



Obrázek 3.2: Mapování okamžíků t_a^i na okamžiky t_s^i při modifikaci melodie s koeficientem $\alpha(t) = 1.5$.

3.1.3 Modifikace hlasitosti

Modifikace hlasitosti se obecně považuje za nejjednodušší. Signál vynásobíme příslušným koeficientem $\gamma(t)$.

$$s'(t) = \gamma(t)s(t) \tag{3.4}$$

Při porovnávání hlasitosti řeči dvou lidí je opět nežádoucí pracovat s tichem. Celkovou obecnou hodnotu energie řečníka získáme zprůměrováním energií dílčích rámců.

3.2 Modifikace spektrální obálky

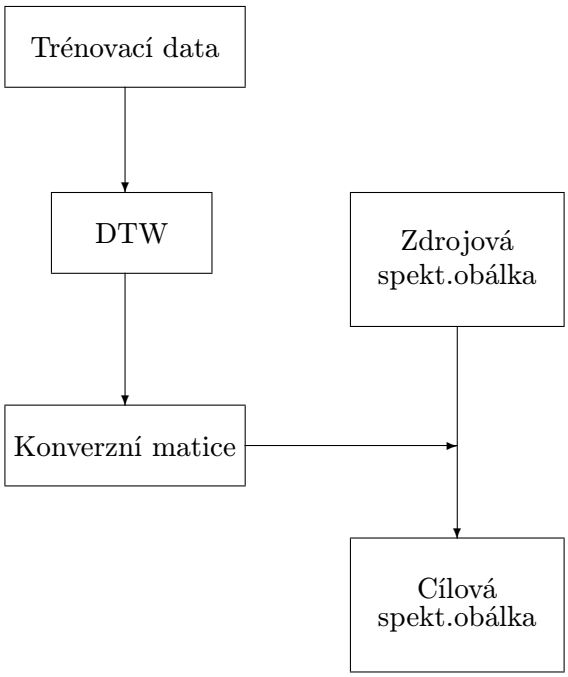
V minulé kapitole jsme si v rámci analýzy řečového signálu pomocí HNM vytvořili obálku amplitud. Tu se nyní budeme snažit modifikovat. Nejprve si vysvětlíme, jak funguje určování parametrů lineární predikce (LPC) a jak jich využít k popsání spektra. Následně si popíšeme metodu dynamického borcení času (DTW), kterou použijeme při zarovnávání dvou promluv na sebe, abychom mohli určit, které části řeči si navzájem přesně odpovídají. Toto je velmi důležité pro náš další krok - trénování *konverzní matice*, s níž provedeme samotnou modifikaci. Pro lepší představu je postup znázorněn na obr. 3.3. Ke zvýšení účinnosti vytvoříme matic několik. Pro znělé úseky, pro neznělé úseky, pro určité třídy fonémů. Rozdělení trénovacích promluv na fonémy provedeme pomocí *fonémového rozpoznávače*.

3.2.1 Kódování pomocí lineární predikce

Artikulační trakt skládající se z hlasivek, hlasového traktu a vyzařování zvuku z úst můžeme popsat následujícím IIR[1] filtrem:

$$H(z) = \frac{1}{1 + \sum_{i=1}^{P} a_i z^{-i}} = \frac{1}{A(z)}, \tag{3.5}$$

[1] Filtr s nekonečnou impulzní odezvou



Obrázek 3.3: Postup při modifikaci spektrální obálky

kde $A(z)$ je polynom P-tého řádu a kde $P = 2k + 1$. Proměnnou k, která představuje počet formantů volíme podle výše hodnoty vzorkovací frekvence. Vzhledem k tomu, že pracujeme s nahrávkami o vzorkovací frekvenci $F_s = 16\,\text{kHz}$, zvolíme počet formantů vyšší.

Koeficienty a_i jsou ovšem neznámé a musí být *identifikovány*. Predikovaný signál $s^{'}(n)$ je popsán jako lineární kombinace předchozích prvků (3.6). Rozdíl mezi skutečnou a predikovanou hodnotou určuje *chybu predikce* (3.7) a samozřejmě, čím menší chyba, tím lepší predikce. Ve výsledku získáme jednoduchý vztah 3.8.

$$s^{'}(n) = -\sum_{i=1}^{P} a_i s(n-i) \tag{3.6}$$

$$e(n) = s(n) - s^{'}(n) \tag{3.7}$$

$$E(z) = S(z)A(z) \tag{3.8}$$

Dalším krokem je minimalizace energie chyby. Vztah 3.9 vyjádříme pomocí známého signálu $s(t)$ a neznámých koeficientů a_i. Samotnou minimalizaci vykonáme za pomoci parciálních derivací (3.10, 3.11).

$$E = \sum_{n} e^2(n) \tag{3.9}$$

$$\frac{\delta}{\delta a_j}\left\{\sum_n [s(n)+\sum_{i=1}^{P} a_i s(n-i)]^2\right\} = 0 \tag{3.10}$$

$$\sum_n s(n)s(n-j)+\sum_{i=1}^{P} a_i \sum_n s(n-i)s(n-j) = 0 \tag{3.11}$$

Po substituci

$$\sum_n s(n-i)s(n-j) = \phi(i,j) \tag{3.12}$$

můžeme výše uvedený vztah přepsat na následující:

$$\sum_{i=1}^{P} a_i \phi(i,j) = -\phi(0,j) \quad \text{pro} \quad i \le j \le P \tag{3.13}$$

Toto už můžeme řešit jako soustavu lineárních rovnic. Výpočet $\phi(.,.)$ lze provést pomocí korelační metody a je popsán např. v [2].

Z LPC koeficientů lze vypočítat LPC-Cepstra(LPCC):

$$c(n) = \mathcal{F}^{-1}\left[\ln\left|\frac{G}{A(z)}\right|^2_{z=e^{j2\pi f}}\right], \tag{3.14}$$

kde G je gain "syntetizačního" filtru.

Nultý LPCC koeficient obsahuje informaci o energii příslušného řečového rámce. Další lze vypočítat pomocí rekurentních vztahů.

LPC-cepstra budou hrát klíčovou roli při zarovnávání trénovacích dat (metoda DTW) a pomohou nám také, až budeme vyhodnocovat kvalitu našeho systému pro konverzi hlasu a porovnávat promluvy cílového řečníka s konvertovanými promluvami zdrojového řečníka. Pomocí LPCC lze totiž snadno vypočítat *logaritmickou spektrální vzdálenost* mezi dvěma řečovými rámci.

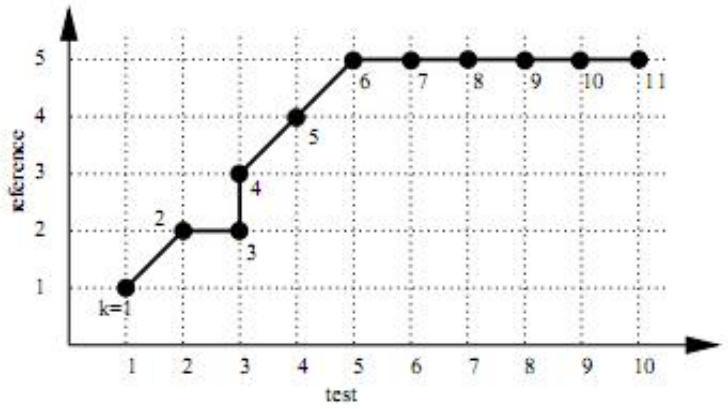
3.2.2 Metoda dynamického borcení času

Dynamické borcení času (DTW) se nejčastěji používá v systémech pro rozpoznávání. V principu totiž na sebe dané vzorky zarovná, vypočte vzdálenost mezi příslušnými vektory (LPCC koeficienty) a ukáže nám, do jaké míry jsou si podobné. My využijeme jeho zarovnávací vlastnosti při *trénovací fázi* tvorby konverzní matice, kdy se budeme snažit získat co největší množství odpovídajících si dvojic vektorů.

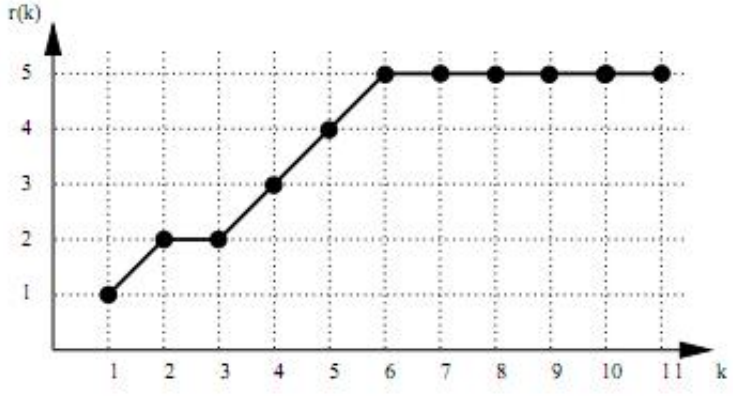
Vždy budeme pracovat s dvěma sekvencemi vektorů. Jednu budeme nazývat zdrojovou a budeme ji značit $\mathbf{R} = [\mathbf{r}(1), \ldots, \mathbf{r}(R)]$, druhou nazveme cílovou, $\mathbf{O} = [\mathbf{o}(1), \ldots, \mathbf{o}(T)]$. Konstanty R a T nám říkají, jak jsou tyto sekvence dlouhé.

Pokud si nadefinujeme obecnou časovou proměnnou k, na jednu osu vyznačíme zdrojovou nahrávku a na druhou osu cílovou nahrávku, můžeme jejich srovnání zakreslit pomocí cesty tak, jak to je ukázáno na obrázku 3.4. Tento a ostatní obrázky z této podkapitoly byly převzaty z [2].

Abychom si naše srovnání mohli pro jednotlivé sekvence nakrokovat, nadefinujeme zdrojovou transformační funkci $\mathrm{r}(k)$ a cílovou transformační funkci $\mathrm{t}(k)$. Jejich průběh vidíme na obrázcích 3.5 a 3.6.



Obrázek 3.4: Cesta pro srovnání zdrojové a cílové nahrávky



Obrázek 3.5: Funkce r(k) pro krokování zdrojové nahrávky.

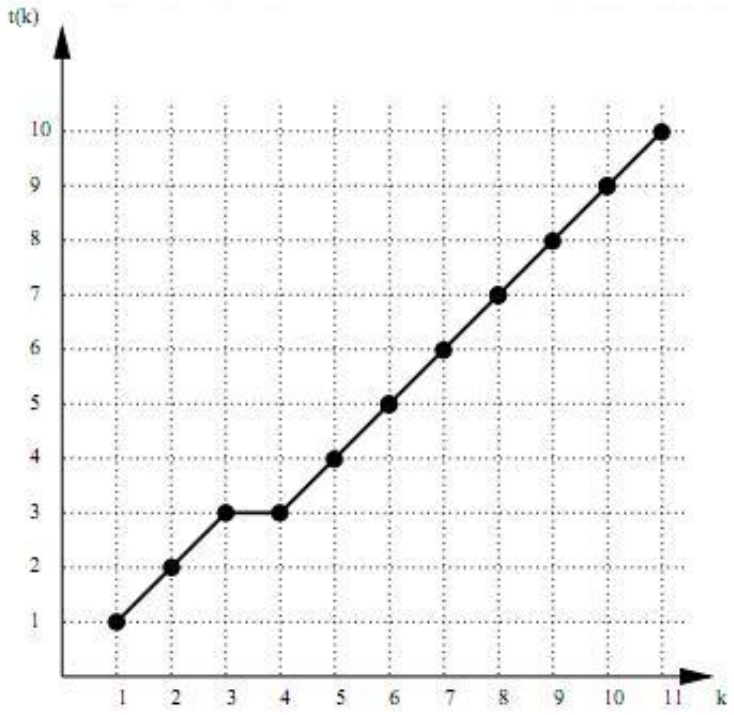
Cesta C je dána svojí délkou K_C a průběhem funkcí $r_C(k)$ a $t_C(k)$.

$$D_C(\mathbf{O},\mathbf{R}) = \frac{\sum_{k=1}^{K_C} d[\mathbf{o}(t_C(k)),\mathbf{r}(r_C(k))]W_C(k)}{N_C}, \tag{3.15}$$

kde $D_C(\mathbf{O},\mathbf{R})$ je vzdálenost mezi sekvencemi **O** a **R** pro právě tuto cestu, $d[\mathbf{o}(.),\mathbf{r}(.)]$ je vzdálenost dvou vektorů, W_C je váha k-tého kroku a N_C je normalizační faktor na váhách závislý. Celková vzdálenost mezi sekvencemi **O** a **R** je pak ta minimální, ze všech možných cest.

Je důležité definovat možnosti průběhů funkcí $r(k)$ a $t(k)$. Určíme si tedy nyní několik pravidel:

- Cesta se nemůže vracet zpět
- Cesta nemůže skákat přes několik vektorů, každý musí být zastoupen alespoň jednou.
- Začátek cesty je v bodech $r(1) = 1$, $t(1) = 1$.

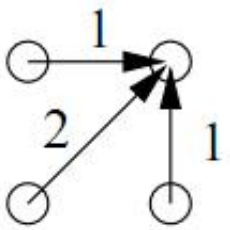


Obrázek 3.6: Funkce t(k) pro krokování cílové nahrávky.

- Konec cesty je v bodech $r(K) = R$, $t(K) = T$.

Nyní je potřeba zvolit váhovací funkci $W(k)$. My použijeme symetrickou (obr 3.7)

$$W_k = [t(k) - (t(k-1))] + [r(k) - (r(k-1))] \tag{3.16}$$



Obrázek 3.7: Symetrická váhovací funkce

a k ní příslušný normalizační vektor

$$N_a = T + R \tag{3.17}$$

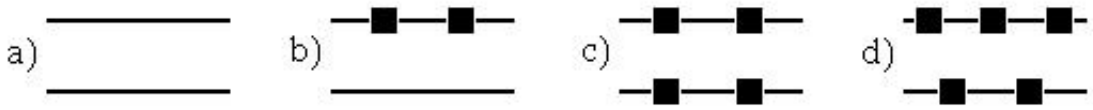
V tuto chvíli máme všechny důležité elementy k efektivnímu výpočtu minimální vzdálenosti D($\mathbf{O}, \mathbf{R}$). Postup je následující:

1. Vytvoříme matici d o velikosti $T \times R$ a vyplníme jí vzdálenostmi mezi příslušnými vektory(každý s každým).
2. Vytvoříme matici g o velikosti $(T+1) \times (R+1)$ s tím, že:
   - $g(0,0) = 0$
   - $g(0, m \neq 0) = g(n \neq 0, 0) = \infty$
3. Určíme si možné předchůdce $g(n,m)$.
   - $g(n, m-1) + d(n,m)$
   - $g(n-1, m-1) + 2d(n,m)$
   - $g(n-1, m) + d(n,m)$
4. Postupně vypočteme částečnou kumulovanou vzdálenost pro jednotlivé body
$$g(m,n) = \min(\text{predchudci})[g(\text{predchudce}) + d(m,n)w(k)] \tag{3.18}$$
5. Kroky 3 a 4 opakujeme pro všechna $n = 1 \ldots T$ a $m = 1 \ldots R$
6. Nakonec vypočteme konečnou minimální normovanou vzdálenost
$$\mathrm{D}(\mathbf{O}, \mathbf{R}) = \frac{1}{N} g(T,R) \tag{3.19}$$

3.2.3 Mapovací funkce

V implementaci HNM, kterou používáme je obálka amplitud zakódována pomocí *diskrétních cepster*. Ty jsou ale přítomny pouze ve znělých rámcích. Tato vlastnost by nám při použití DTW působila velké problémy, protože, jak jsme si řekli v minulé kapitole, při této metodě není možné mezi vektory "skákat". Vytvoříme si tedy funkci, která nám znělé okamžiky hlasivkové periody(t^i) namapuje na náležitá LPC-cepstra a přiřadí k sobě příslušná diskrétní cepstra. Postup je následující:

1. Vypočteme LPC-cepstra pro obě promluvy, které chceme zarovnávat.
2. Použijeme metodu DTW a získáme tak dvousloupcové pole (DTW cestu) hodnot rámců, které si přísluší.
3. Obě promluvy analyzujeme pomocí HNM a vyextrahujeme z nich diskrétní cepstra.
4. Zjistíme, v kterých rámcích se jednotlivá cepstra nacházejí a poznamenáme si to.
5. Postupně procházíme DTW cestu získanou v druhém kroku a zároveň přihlížíme k poznámkám ze čtvrtého kroku. Pro každou dvojici rámců bude platit právě jedna z následujících možností (přehledně je zobrazuje obrázek 3.8):
   a) Žádný rámec neobsahuje diskrétní cepstrum - Pokračujeme dál.
   b) Pouze jeden rámec obsahuje diskrétní cepstrum - Pokračujeme dál.
   c) Oba rámce obsahují právě stejné množství diskrétních cepster - Jednotlivé dvojice si poznamenáme a pokračujeme dál.
   d) Rámce obsahují rozdílné množství diskrétních cepster - Vytváříme dvojice diskrétních cepster, dokud jsou k dispozici v obou rámcích. Zbývající nadbytečná cepstra napárujeme s již použitými z druhého rámce.
6. Výsledkem tedy bude pole příslušících si dvojic diskrétních cepster.



Obrázek 3.8: Případy, které mohou nastat ve 4. kroku mapovací funkce. Černé čtverce symbolizují diskrétní cepstra obsažená v rámcích.

3.2.4 Konverzní matice

Nezbytným krokem k tvorbě konverzní matice je *tréninková fáze*. Vybereme dva řečníky, mezi kterými chceme provádět konverzi hlasu. Nahrajeme velké množství krátkých nahrávek, kde oba říkají stejné, předem připravené sekvence slov. Z každé dvojice nahrávek vyextrahujeme pole dvojic odpovídajících si diskrétních cepster tak, jak jsme si popsali v minulé podkapitole. Tato pole mezi sebou konkatenujeme a získáme tak jeden větší set dvojic vektorů. Finální zpřesňovací úprava nad trénovacími daty bude odečtení průměrných vektorů μ_z a μ_c. Vzhledem k pravidlům násobení matic nemůžeme zdrojové vektory s těmi cílovými jednoduše vynásobit. Stejně tak nelze na jednu z matic použít obyčejnou inverzi, protože matice nejsou regulární. Provedeme Moore-Penrose pseudoinverzi. Je nutno poznamenat, že tento krok je vždy zatížen nějakou chybou.

V tuto chvíli máme vytvořenou čtvercovou konverzní matici o velikosti počtu koeficientů diskrétních cepster. Tu si uchováme a kdykoliv budeme provádět konverzi ze zdrojového mluvčího na cílového, použijeme ji. Ve fázi syntézy hlasu zdrojového mluvčího vynásobíme touto maticí postupně všechna jeho diskrétní cepstra, tak jak ukazuje následující rovnice:

$$\mathbf{c}_c = [\mathbf{M}_T(\mathbf{c}_z - \mu_z)] + \mu_c, \tag{3.20}$$

kde c_z je zdrojové cepstrum, c_c cílové cepstrum, μ_z zdrojový průměrný vektor, μ_c cílový průměrný vektor a $\mathbf{M}_T$ je transformační (konverzní) matice.

Nyní už nepotřebujeme žádné nahrávky cílového mluvčího. Konverzní matice samozřejmě není univerzální a pro každou dvojici mluvčích je potřeba vytvořit novou (Případně dokonce dvě, pokud chceme konvertovat oběma směry.).

3.2.5 Rozšíření počtu konverzních matic

Abychom zvýšili kvalitu konverze, dekomponujeme si promluvu na několik úseků, pro které budeme vytvářet samostatné konverzní matice. Kromě znělých částí signálů to budou i neznělé a použijeme i rozdělení fonémové. Rozdělení promluv na jednotlivé části pro nás samozřejmě bude znamenat nutnost obstarání si větší množiny nahrávek. Počet trénovacích vzorků, určených pro jednotlivé typy matic, extrahovatelných z jedné dvojice nahrávek bude logicky nižší.

Konverzní matice pro šumovou složku

U šumové složky se nemusíme zabývat mapováním okamžiků hlasivkové periody na rámce a můžeme použít DTW srovnání pomocí LPC-Cepster. Kvůli absenci základního tónu je zde perioda zvolena pevně a přesně kopíruje rámce. Opět tedy projdeme trénovací fází stejně

jako u harmonické složky s tím rozdílem, že namísto diskrétních cepster budeme používat LPC-Cepstra. Tyto také budeme ve výsledku(Ve fázi syntézy) modifikovat vytvořenou maticí určenou pro šumovou složku.

Budeme také rozlišovat znělé rámce obsahující jak harmonickou, tak i šumovou složku a rámce neznělé, pouze šumové. Matice pro šumovou složku budou tudíž dvě.

Konverzní matice pro fonémové třídy

Další prostředek ke zkvalitnění konvertovaného řečového výstupu, který použijeme, bude *fonémový rozpoznávač* vyvíjený skupinou Speech@FIT. Ten zanalyzuje řečový signál a s relativně nízkou chybovostí [3] vytyčí časové hranice trvání jednotlivých fonémů. Dohromady dokáže rozeznávat celkem 45 českých a 39 anglických fonémů. S nahrávkami v jiných jazycích pracovat nebudeme. Fonémy můžeme rozdělit na několik tříd. Nás budou zajímat neokrouhlé (a,e) a okrouhlé (o,u) samohlásky. Širší dělení by mohlo přinést ještě lepší výsledky a je jedním z možností výběru směru dalšího vývoje.

Fonémovému rozdělení bude podléhat jak trénovací, tak modifikační etapa. Ve výsledku budou tedy i pro harmonickou složku dvě konverzní matice.

Kapitola 4

Implementace systému pro konverzi hlasu

Připravili jsme si teoretický základ k vytvoření efektivního systému pro konverzi řeči a nyní se pokusíme takový systém implementovat. Některé součásti systému byly se svolením autorů převzaty a více či méně modifikovány. Toto bude vždy explicitně uvedeno.

4.1 Matlab

Rozhodl jsem se pro implementaci ve výpočetním prostředí Matlabu, který má následující vlastnosti:

Výhody

- Jednoduchost a uživatelská přístupnost
- Zobrazovací funkce
- Rozsáhlá a přehledná nápověda
- Automatická inicializace proměnných
- Snadná práce s maticemi(signály)

Nevýhody

- Výkon
- Nízká podpora objektově orientovaného přístupu

Celou analytickou a trénovací fázi, které zahrnují převážnou část operací nad daty, budeme provádět předem, proto nás nízký výkon Matlabu nebude nijak výrazně ovlivňovat. Významným faktorem k tomu, abych zvolil toto výpočetní prostředí byla skutečnost, že ho použil i autor implementace HNM, která je jádrem našeho systému pro konverzi hlasu. Celkově výhody výrazně převažují nad nevýhodami.

4.2 Fáze analýzy

K analýze použijeme HNM implementovaný I. Szökem. Konkrétně funkci `ParametrizeFile`, která si sama volá další, podstatné k parametrizaci. V ní inicializujeme počáteční hodnoty konstant, jako je vzorkovací frekvence, minimální a maximální hodnota základního tónu, koeficienty pro jeho vyhledávání, počet koeficientů vektorů diskrétních cepster, délka a posun rámců signálu. Parametrizační funkci obohatíme o výpočet LPC-Cepster. Počet jejich koeficientů přizpůsobíme počtu koeficientů diskrétních cepster. Dále přidáme výpočet toho, které okamžiky hlasivkových period (t^i) patří ke kterým rámcům.

Zarovnávání promluv zajistí funkce `dtw`, jež byla představena v laboratořích předmětu ZRE[1]. Vstupními argumenty jsou matice LPC-cepster.

Než přistoupíme k samotnému tréninku konverzních matic, musíme z trénovacích dat získat pomocí DTW zarovnaná jak LPC-Cepstra, tak i diskrétní cepstra. Pro tento účel byla implementována mapovací funkce `DTWPathDceps`, popsaná v sekci 3.2.3, jež si za argument bere LPCC DTW cestu a pomoci ní a informace, ve kterých rámcích se nachází jednotlivá dískrétní cepstra, vypočítá jejich DTW srovnání.

Abychom mohli pracovat s diskrétními cepstry, musíme za pomoci funkce `GetDCeps` spustit falešnou syntézu, která se na rozdíl od té reálné ukončí ihned po jejich výpočtu, který zajistí funkce `GetNewAkFromAmplitudeEnvelope` modifikovaná tak, aby kromě svých standardních výpočtů navíc ještě vracela jednotlivé vektory diskrétních cepster. Ty si postupně ukládáme do matice.

4.3 Tvorba konverzních matic

4.3.1 Harmonická složka

Pro harmonickou složku tvoříme dvě konverzní matice. Pro okrouhlé a neokrouhlé samohlásky. Použijeme fonémový rozpoznávač, který jsme si opatřili na oficiálních stránkách Speech@FIT[2]. Po jeho použití dostaneme přesné časy trvání jednotlivých fonémů. Při získávání trénovacích dat budeme tato dělit na příslušné skupiny. Postup pro tvorbu obou matic je identický. Rozdílná jsou pouze trénovací data.

Každý vektor diskrétního cepstra v sobě na první pozici nese informace o energii, která je pro trénink matic irelevantní. Proto je pomocí `ObtainTrainData` ořízneme. Tato funkce se také postará o to, aby se trénovací data dostala do konečného formátu určeného k tréninku. Prochází DTW cestu diskrétních cepster a rovná je do příslušných matic tak, aby byly k sobě patřící vždy na stejných pozicích. Tvorbu konverzní matice obstarává `MakeModMatrix`. O Moore-Penrose pseudoinverzi se postará zabudovaná funkce Matlabu `pinv`.

4.3.2 Šumová složka

Tvorba konverzních matic pro šumovou složku se bude od tvorby matic pro harmonickou složku lišit pouze ve sběru trénovacích dat. Zatímco u harmonické složky jsme shromažďovali diskrétní cepstra vyskytující se pouze ve znělých rámcích, zde budeme sbírat LPC-Cepstra jak ze znělých (první konverzní matice), tak i z neznělých (druhá konverzní matice) rámců. Mapovací funkci používat nebudeme.

[1] Zpracování řečových signálů - Povinný předmět magisterského oboru Počítačová grafika a multimédia na FIT VUT v Brně.

[2] http://speech.fit.vutbr.cz/

4.4 Fáze syntézy a konverze

Syntéza je též součástí HNM a má ji na starost funkce `SyntheseFile`. Na začátku je opět možnost přizpůsobit si hodnoty konstant našim potřebám. Průběh syntézy je popsán v kapitole věnované HNM. Modifikace prozodie bude spočívat ve správném nastavení koeficientů $\alpha(t), \beta(t)$ a $\gamma(t)$. To zajišťují funkce `PitchMod`, `TSMod`, `IntensityMod`, případně, pro všechny tři najednou, obecná `Mod`. Jejich průběh spočívá v nalezení průměrných hodnot daných atributů, jejich porovnání mezi sebou a podle výsledku nastavení příslušného koeficientu. Pro modifikaci spektrální obálky, nahradíme funkci `GetNewAkFromAmplitudeEnvelope` funkcí `DCepsModification`. Uvnitř se bude odehrávat násobení diskrétních cepster náležitými konverzními maticemi. Modifikaci šumové složky (násobení LPC-Cepster) provedeme přímo ve funkci `SyntheseFile`.

V rámci experimentů s diskrétními cepstry byly implementovány funkce `SubstituteDCeps` a `GetNewAkChangeDCeps`, které z cílové nahrávky získají diskrétní cepstra a nahradí jimi ta ze zdrojové nahrávky. Tato metoda sice přináší velmi uspokojivé výsledky, ale vzhledem k nutnosti použití nahrávky cílového mluvčího není v praxi příliš použitelná.

Kapitola 5

Testování a výsledky

Testovány byly jak jednotlivé části systému zvlášť, abychom viděli významnost dílčích prvků, tak i systém jako celek. Proběhly tudíž testy pro modifikaci rytmu, základního tónu, hlasitosti, modifikaci spektrální obálky pomocí jedné matice(harmonická složka), modifikaci spektrální obálky pomocí širšího spektra konverzních matic, modifikaci spektrální obálky záměnou diskrétních cepster a nakonec celková konverze(Modifikace prozodie a modifikace spektrální obálky pomocí širšího spektra konverzních matic).

5.1 Testovací data

Prozodické modifikace byly prováděny na TIMIT[1] korpusu, který podle [7] obsahuje nahrávky 630 řečníků z osmi největších dialektů americké angličtiny, od každého po deseti větách. Nahrávky jsou k dispozici ve vzorkovacích frekvencích 8 kHz a 16 kHz kódované na 16 bitů a jsou zhruba 4 s dlouhé. TIMIT korpus byl vytvořen přímo pro účely výzkumu v oblasti zpracování řeči. My použijeme, jak jsme si už dříve řekli, nahrávky s $F_s = 16$ kHz.

V rámci vytváření TIMIT korpusu bylo sice každým řečníkem namluveno 10 řečových nahrávek, ale pouze dvě z nich obsahuji slovní sekvence stejné pro všechny řečníky. Tato množina je pro trénink konverzních matic nedostatečná. Proto jsme nuceni si nahrát vlastní trénovací set. Při nahrávání byly použity stejné postupy jako u TIMIT korpusu (16 bitů, 16 kHz). Pro každého mluvčího bylo nahráno celkem dvacet řečových signálů. Průměrná délka jedné nahrávky je asi 3 s.

5.2 Výsledky

Všechny modifikované nahrávky jsou zatíženy zkreslením, způsobeném harmonickým a šumovým modelem. Je tedy vhodné originální nahrávky analyzovat a vzápětí syntetizovat beze změny jakýchkoliv parametrů a pro hodnocení úspěšnosti systému konverze pak modifikované nahrávky porovnávat s nimi. Výsledky jsou k dispozici na přiloženém CD rozdělené do adresářů podle typu konverze.

5.2.1 Prozodické vlastnosti

Modifikace rytmu dopadly podle očekávání výborně. Nahrávka byla natažena/smrštěna podle cílového řečníka bez slyšitelné ztráty na kvalitě. Stejně tak celková hlasitost. Obecná

[1] Acoustic-Phonetic Continuous Speech Corpus

metoda pracující s průměrným základním tónem získaným z celých nahrávek ho sice pro zdrojového mluvčího nezmění na přesnou hodnotu cílového mluvčího, ale posun správným směrem je zřetelně slyšitelný. Při modifikaci základního tónu směrem k řečníkovi se základním tónem výrazně nižším byl zaznamenán pokles kvality projevující se občasným "přeskočením" hlasu. Jak jsme se dozvěděli již v [6] a [5], modifikace základního tónu má také svoje hranice. Za těmi pak kvalita syntetizovaného signálu prudce klesá.

Člověk má tendenci prozodické parametry v čase měnit. Důležité informace sděluje pomaleji, hlasitěji, nepodstatné pak rychleji, tišeji. Třídy fonémů nebo dokonce jednotlivé fonémy mají charakteristické hodnoty těchto parametrů pohybující se v určitých intervalech. V dalším vývoji by bylo proto vhodné nahrávku rozdělit (např. fonémovým rozpoznávačem) a výpočty a změny náležitých parametrů provádět lokálně.

5.2.2 Spektrální obálka

Modifikace spektrální obálky pomocí jedné konverzní matice (harmonická složka) nebyla tak účinná, jak se předpokládalo. Její aplikace na konverzi hlasu dvou mužských hlasů byla jen stěží postřehnutelná. Jisté zlepšení nastalo se zakomponováním odečítání a přičítání průměrných cepstrálních vektorů. Abychom lépe slyšeli účinek této modifikace, byla provedena konverze mezi pohlavími.

Dekompozice trénovacích dat a modifikace pomocí rozšířeného spektra konverzních matic výrazně přispěla k posunu hlasu zdrojového mluvčího směrem k hlasu cílového mluvčího. Stále ale není výsledek takový, aby ho bylo možné zaměnit za hlas z originálních nahrávek cíle. Zdokonalení bychom mohli dosáhnout rozšířením množiny trénovacích dat nebo jejich dalším štěpením.

Jsou k dispozici i výsledky experimentální metody nahrazení zdrojových diskrétních cepster cílovými. Díky použití originálních cílových parametrů je výsledek modifikace v některých částech téměř nerozeznatelný od originálu. Problém nastává při přechodu mezi znělými a neznělými rámci, kde hlas skáče ze zdrojového na cílového mluvčího. Tento experiment je ovšem z hlediska konverze hlasu slepou uličkou, protože našim cílem je ve fázi syntézy pracovat pouze se zdrojovými daty. Identickou promluvu namluvenou oběma řečníky v drtivé většině případů mít nebudeme.

5.2.3 Celková konverze

Kompletní konverze je samozřejmě kvalitnější než ty jednotlivé, dílčí. Jejich samostatná aplikace nám ale umožňuje v modifikované nahrávce slyšet jejich podíl a my tak můžeme ručně doladit jednotlivé parametry. V tomto stádiu vývoje bývá ruční konverze kvalitnější než obecná.

Kapitola 6

Závěr

Obsahem bakalářské práce je prostudování metod konverze hlasu a jejich následné uplatnění.

V úvodu jsme si nastínili, jaké informace může obsahovat lidský hlas. Objasnili jsme, co to je konverze hlasu a z jakých fází se skládá. Ustanovili jsme, které metody a nástroje budeme používat a řekli si, kde k nim můžeme najít relevantní informace.

Druhá kapitola popisuje harmonický a šumový model, který slouží k parametrizaci a syntéze nahrávek, tak, jak byl implementován v [6].

V třetí kapitole jsou diskutovány metody konverze hlasu. Nejprve se zabývá modifikacemi prozodie vycházejícími z HNM a poté možnostmi alternace spektrální obálky. Je zde představena metoda pracující s konverzními maticemi a vysvětleno použití LPC a DTW k jejich tréninku. Je nastíněna možnost rozšíření jejich počtu. Zabývá se také řešením problému nalezení odpovídajících si diskrétních cepster.

Obsahem čtvrté kapitoly je postup implementace jednotlivých prvků systému pro konverzi hlasu. Ukazuje výhody a nevýhody Matlabu a důvody proč bylo toto výpočetní prostředí zvoleno. Je objasněno, které části implementace byly převzaty a odkud.

Poslední část bakalářské práce se věnuje průběhu testů a zamýšlí se nad dosaženými výsledky. Naznačuje, co by mohlo vést k jejich zlepšení. Zmiňuje se také o tom, s jakými nahrávkami jsme pracovali.

Konverze hlasu souvisí s velkým množstvím vědních oborů od biologie přes zpracování signálů, lingvistiky, až třeba po pravděpodobnost. Existuje tedy mnoho směrů, kterými se v dalším vývoji můžeme ubírat. Pokud chceme pokračovat v používání metod popsaných v této bakalářské práci, bylo by vhodné zvýšit množinu trénovacích dat, provést jejich širší dekompozici a natrénovat více typů konverzních matic. Zkvalitnění konverze by také přineslo širší uplatnění fonémového rozpoznávače a to zejména v oblasti modifikace prozodických parametrů. Další možností je pokusit se zlepšit implementaci HNM, případně tento model vyměnit za jiný.

Abychom mohli zlepšit kvalitu konverze různými pomocnými algoritmy, je potřeba lépe porozumět tomu jakým způsobem se řeč vytváří a jak ji lidé celkově vnímají. Podle [1] například, pokud sedíme v kavárně a chceme zvýšit hlas, nezvyšujeme hlasitost celého signálu, ale pouze jeho části. Převážně u souhlásek. Při myšlení v širších souvislostech a zvažování různých faktorů ovlivňující lidský hlas se pak ani modifikace hlasitosti nezdá tak snadná, jak se může na první pohled zdát.

Literatura

[1] Benesty, J.: *Springer Handbook of Speech Processing.* Springer, Berlin, 2008.

[2] Cernocky, J.: *Studijní opora: Zpracování řečových signálů.* FIT VUT v Brně, 2006.

[3] Matejka, P.; Schwarz, P.; Cernocky, J.; aj.: *Phonotactic Language Identification using High Quality Phoneme Recognition.* in Proc. Eurospeech2005, 2005.

[4] Schwarz, P.: *Phoneme Recognition based on Long Temporal Context.* PhD Thesis, University of Technology, Brno, 1999.

[5] Stylianou, Y.: *Harmonic plus noise model for speech, combined with statistical methods, for speech and speaker modification.* PhD Thesis, ENST, Paris, 1996.

[6] Szoke, I.: *Prozódie syntetické řeči.* DP, FIT VUT v Brně, 2003.

[7] WWW stránky: TIMIT Acoustic-Phonetic Continuous Speech Corpus. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1`.